Bounds on the Mean Classification Error Rate of Multiple Experts

Padhraic Smyth
Jet Propulsion Laboratory, MS 5253660
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109
email:pjs@aig.jpl.nasa.gov

September 22, 1995

Abstract

A database contains N items, each item belonging to one and only one of a finite set of classes. The *true* class labels for these items are unknown. K experts each provide a set of N classification labels for the N items in the database. In this paper it is shown that given the experts' labels, one can compute simple bounds on the average classification accuracy of the experts relative to the unknown true labels. No assumptions are made about the labelling patterns of the experts or the nature of the data. The bounds are useful in practical classification problems where absolute ground truth is unknown and experts must subjectively provide labels for feature data. The method is applied to the problem of assessing the collective accuracy of geologists who count volcanoes in images of Venus.

Keywords:

classification, multiple experts, remote-sensing, volcano catalog

1 Introduction and Notation

Consider that a person (an observer) has a database of N items, each described by a feature vector $\underline{x}_i, 1 \leq i \leq N$. Each item belongs to one of m classes, $m \geq 2$: the classes are mutually exclusive and exhaustive. It is assumed that for each item $\underline{x}_i$ there exists a true label ω_i (a reference label) which is unknown. For example, if the $\underline{x}_i$ were pixel measurements of an object of unknown class in a remotely-sensed image, the true class label could in principle be obtained by visiting the ground site and ascertaining the class of the object in an unambiguous manner (so called "ground truth").

The observer is assumed to have no information whatsoever about the true class labels of the items. Let K experts (K > 1) each provide a set of N labels for the N items, i.e., each expert examines each item $\underline{x}_i$ in turn and provides a subjective estimate of the true class label for that item. Define $\bar{e}$ as *the mean classification error rate, averaged across the K experts, relative to the true labels*, i.e., over all the experts, a certain fraction of items have been mislabelled relative to the truth. By definition

$$\bar{e} = \frac{1}{KN} \sum_{k=1}^{K} \sum_{i=1}^{N} e_{ik} \tag{1}$$

where, for the label of labeller k on the ith item, $e_{ik} = 1$ if it is in error and $e_{ik} = 0$ if it is correct.

The fact that $\bar{e}$ is defined as the *mean* error rate of K labellers rather than the error rate of any one labeller is a key point and enables calculation of the bounds. Without knowing ground truth one can not make any statements about the errors of an individual labeller. References to "errors" will be assumed to mean "errors relative to ground truth" throughout the paper.

2 Motivation and Background for this Problem

Assessing the collective classification accuracy of a group of experts on a database is an important issue in certain practical classification problems. For example, scientists subjectively label pixels or regions in a remote-sensing image into a set of known ground-cover classes, or medical specialists classify medical records into particular diagnostic classes. In such cases obtaining the true class labels for the data is frequently either physically impossible or prohibitively expensive. For example, in remote-sensing it may be impractical to visit the remote sites to ascertain ground truth. In medical diagnosis it may be too expensive to perform the necessary tests or surgery to determine with absolute certainty what disease the patient actually had. in classification-oriented applications, as online data become more readily available, the proportion of the data for which the true class labels are known is likely to continue to decrease. Thus, being able to infer statements about the accuracy of human experts is quite valuable in these types of problems. In Section 4 we describe a particular application of the method to counting volcanoes in radar images of Venus. The volcano counting problem originally motivated this work: it is a problem of considerable geologic importance involving multiple expert opinions.

3 A Lower Bound on $\bar{e}$

From Equation (1), the average error rate can be written as

$$\bar{e} = \frac{1}{KN} \sum_{i=1}^{N} e_i \tag{2}$$

where $e_i = \sum_{k=1}^{K} e_{ik}$, is the total number of errors made on item i, $0 \leq e_i \leq K$.

Consider the ith item. Let n_{ij} be the number of times that label j was provided by the K labellers for item i, $0 \leq n_{ij} \leq K$.

Let j^* indidate the correct label for the item. Thus $K - n_{ij^*}$ is the number of errors made on the ith item. Since j^* is unknown, one has

$$e_i \geq \min_j \{K - n_{ij}\}, \qquad 1 \leq j \leq M. \tag{3}$$

Thus,

$$\bar{e} \geq \frac{1}{KN}\sum_{i=1}^{N}\min_{j}\{K - n_{ij}\} = \frac{1}{KN}\sum_{i=1}^{N}\left(K - \max_{j}\{n_{ij}\}\right). \tag{4}$$

This bound is a function of the number of disagreements made by the labellers. If there are no disagreements, the bound is 0. In the worst-case scenario the labellers agree on all items but are incorrect in each case, yielding a lower bound of 0 while the true error rate is 1. In general, however, the bound will be non-zero for practical problems, thus providing an indication of the overall error rate of a set of experts. Note that at least one of the K labellers must have an error-rate *greater than or equal to* the lower bound: thus, for example, even if the labellers are noted experts, the bound will imply that at least one of them has an error rate greater than some value, relative to ground truth. If this value is large (say greater than 10%) it may indicate the need to re-evaluate the quality of the feature data $\underline{x}_i$, or the quality of the expert labelling process, or both.

Equation (4) is the lowest bound one can obtain on the mean error rate without additional information about the problem being available. For example, if K = 2 and one of the labellers is always correct, then the bound is exactly the mean error rate.

3.1 Binary Classification

With binary classification, $m = 2$, we can index the labelling patterns by the number of labels belonging to one of the classes ("detections"), $0 \leq d \leq K$. Let n_d be the number of labelling patterns which have d detections ($\sum_{d=0}^{K} n_d = N$ if all items are labelled). For example, n_1 is the number of items each of which were labelled as a detection by only one of the K labellers. For binary labels, the bound reduces to:

$$\bar{e} \geq \frac{1}{KN}\sum_{d=0}^{K} n_d\left(K - \max_{j}\{n_{ij}\}\right) \tag{5}$$

$$= \frac{1}{KN}\sum_{d=0}^{K} n_d \min\{K - d, d\} = \frac{1}{KN}\sum_{d=1}^{K-1} n_d \min\{K - d, d\} \tag{6}$$

3.2 Binary Classification with Two Labellers

With K = 2,

$$\bar{e} \geq \frac{n_1}{2N} \tag{7}$$

where n_1 is the number of items labelled by the 2 experts where they disagree, i.e., one gets the simple result that the mean error rate is lower bounded by half the fraction of disagreements. If two labellers disagree on all items, their mean error rate must be 0.5 (which also equals the bound in this case).

3.3 An Upper Bound on $\bar{e}$

One can also derive a simple upper bound on $\bar{e}$:

$$\bar{e} \leq \frac{1}{KN}\sum_{i=1}^{N}\max_{j}\{K - n_{ij}\} \tag{8}$$

$$= \frac{1}{KN}\sum_{i=1}^{N}\Big(K - \min_{j}\{n_{ij}\}\Big). \tag{9}$$

This upper bound is always greater than or equal to $(1 - \frac{1}{m})$. Thus, it is of limited value in practice, since it says that the mean error rate per labeller is no worse than $1 - \frac{1}{m}$, which in turn, for reasonably-sized m, is quite close to 1 (the trivial upper bound).

4 Application of the Lower Bound

4.1 Catalog Generation in Scientific Applications

In a number of observational sciences such as astronomy and planetary geology, a common step in the scientific process is to convert raw data (such as images) into a catalog of objects of interest (Fayyad et al. (in press)). Such catalogs form a standard data product which can be used by other scientists as the basis for quantitative scientific studies (such as investigations of the spatial clustering patterns of objects, etc.). Examples include counting stars and galaxies in telescope images to generate a sky catalog, counting impact craters on the surface of the moon, counting and characterizing sunspots in images of the Sun, and counting volcanoes in radar images of Venus. Typically the cataloging is carried out by known experts in the field.

In each of these applications, the quality of the final catalog is inevitably a function of the subjective nature of the cataloging process. In some applications there may be little variation between the labels provided by different experts for the same object: in other applications the variance may be quite high, indicating that the data in the catalog should be treated accordingly. The variation in expert opinion may be due to visual ambiguity introduced by the resolution limits of the data, perhaps the pixel-resolution of an imaging instrument. The lower bounds on error rate described earlier provide a simple method to ascertain a lower bound on the accuracy of the subjective cataloging process: an application to volcano counting is described below.

4.2 Bounding the Mean Accuracy of Volcano Counting

The Magellan spacecraft orbited Venus from 1990 to 1994 and transmitted back to Earth a high resolution synthetic aperture image map of the planet, approximately 30,000 1Mybte images in total. The study of volcanic features on the surface of Venus is a key issue in planetary geology due to the predominance of volcanism on the planet (Saunders et al. (1992)). Generating a comprehensive volcano catalog from the Magellan data is a prerequisite for more advanced studies such as cluster analysis of the volcano locations. Of interest in the context of this paper is the accuracy of the volcano labels provided by planetary geologists.

In previous work a pattern recognition system for automatically counting volcanoes in the Magellan images of Venus has been developed: the pattern recognition system is described in detail elsewhere (Burl et al. (1994a)) and is not of direct interest here. As part of the development of the pattern recognition system, several planetary geologists, considered experts in Venus volcanism, provided labels for sets of Venus images as training and test data. Significant variability between the geologist's labellings was noticed, thus motivating work on the problem of quantifying classification accuracy of both humans and algorithms in the absence of ground truth. The variability in the labelling appears to be primarily due to the relatively low signal-to-noise ratio (relative to small volcano structure) in the SAR images (Fayyad et al. (in press)).

Each geologist examined sets of images independently and used lnc)use-clicks within a graphical user-interface to indicate their estimate of where the volcanoes were located within a given image. The first labelling experiment consisted of 4 images and 4 experts (geologists A, B, C, and D). Between the 4 geologists, 269 estimated volcano locations were found in total in the 4 images. Consider this to be the database of $N = 269$ items with binary labels: volcano or non-volcano. One can think of each "item" as a local pixel window or region of interest. The lower bound on mean error rate (using Equation (4)) was found to be 19.370, i.e., the average error rate among geologists A, B, C, and D, labelling volcanoes on these particular 4 Magellan images, is at least 19.3% relative to ground truth' .

The second labelling experiment consisted of 2 geologists (A and B from the first experiment) who each individually labelled 38 images (different from the first 4). In this case 512 possible volcano locations were found in total. Again, considering this to be a database of $N = 512$ items with binary labels results in a lower bound on the mean error rate of A and B of 24 .1%. If only the labellings of geologists A and B arc! considered on the 4 images in the first experiment, they made at least 22.2'%0 errors on average (for these 2 geologists on these 4 images).

Across different subsets of images, with different sets of geologists, the results for the volcano problem have consistent] y shown a lower-bound on the mean error rate of about 20%. Thus, one can state that typically at least one of the expert geologists is in error at least 20% of the time in terms of volcano labelling, over a range of different Magellan images, relative to the ground truth. The true mean error rate for the geologists could in fact be much higher than 20%: the fact that it is at least this high indicates that interpretation of volcano catalogs obtained from the Magellan dataset (even those provided by the experts in the field) requires some care. In terms of developing a pattern recognition system, the result shows that both the training and test data sets contain a significant degree of noise in the labels and this must be taken into account both in training and evaluating any such system (Burl et al. (1994b), Smyth et al. (in press)).

5 Comments on Related Work

Previous work on modelling noise in class labels has largely relied on parametric models of the noisy labelling process. For example, *rating models* assume that a set of labellers provides a discrete set of ratings of the likelihood that an item belongs to a class, and from the ratings of multiple labellers an overall combination model and posterior estimates for individual items are found (French

(1 985), Agresti (1 992), Uebersax (1 993)). A key issue is the nature of the assumptions about the independence of tile different labellers.

In another distinct approach, the error patterns in class labelling arc assumed to obey a particular model and the implications arc analysed (Aitchison and Begg (1 976), Titterington (1989), Lugosi (1 992)). For example, as the noise in the labelling process increases the effect on the estimation performance of certain parametric classification methods has been investigated (Krishnan aud Nandy (1 990)).

Both of these general approaches bear some relation to the problem discussed in this paper aud indeed the ratings approach has been used with success on the volcano data (Smyth et al., in press). However, the results in this paper arc distinct from this prior work in the sense that the bounds derived here make no assumptions whatsoever about the nature of the labelling errors, the independence relationships between the K labellers, or the underlying distribution of the data.

6 conclusion

Simple bounds were derived for the mean classification error rate of K labellers in the absence of ground truth. The lower bound was applied to data from a remote-sensing image analysis problem and the results confirmed that the subjective error rate for the problem is quite high. The method has applications to classification problems where data must be labelled in a subjective manner by experts and there is no ground truth available to calibrate their performance.

Acknowledgements

Thr author gratefully acknowledges useful discussions on this topic with Michael Burl, Usama Fayyad, Pietro Perona, Jayne Aubele, aud Larry Crumpler. The research described in this paper has been carried out by the Jet Propulsion Laboratory, California Institute of '1'ethnology, under contract with the National Aeronautics and Space Administration.

References

Agresti, A. (1992). Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research*, vol.], p~).201-218.

Aitchison, J. aud C. B. Begg (1976). Statistical diagnosis when basic cases arc not classified with certainty. *Biometrika*, 63 (1), pp.1-12.

Burl, M. C., U. M. Fayyad, P. Perona, P. Smyth, and M. P. Burl (1994a). Automating the hunt for volcanoes on Venus. In *Proceedings of the 1994 Computer Vision and Pattern Recognition Conference, CVPR-94*, Los Alamitos, CA: IEEE Computer Society Press, pp.302- 309.

Burl, M. C., U. M. Fayyad, P. Perona, and P. Smyth (1994 b). Automated analysis of radar imagery of Venus: handling lack of ground truth. In *Proceedings of th c IEEE Conference on Image*

Processing, Piscataway, NJ: IEEE Press, vol.]]], pp.236 240.

Fayyad, U. M., P. Smyth, M. C. Burl, and P. Perona (in press). Learning to catalog science images. In *Early Visual Learning*, S. Nayar and T. Poggio (eds.), New York, NY: Oxford University Press.

French, S. (1 985). Group consensus probability distributions: a critical survey. In *Bayesian Statistics 2,* J. M. Bernardo, M. H. DeGroot, D. V. Lindley, A. F. M. Smith (eds.)), Amsterdam, North-Holland: Elsevier Science Publishers, pp.1 83-20'2.

Krishnan, T. and S. C. Nandy (1990). Efficiency of discriminant analysis when initial samples arc classified stochastically. *Pattern* Recognition, vol.23, no.5, pp.529 537.

Lugosi, G (1992). Learning with an unreliable teacher. *Pattern Recognition*, vol. **25,** no.1, pp.79 **87.**

Saunders, R. S., et al. (1992). Magellan mission summary. *Journal of Geophysical Research*, vol .97, no.E8, pp.] 3067-13090.

Smyth, P., M. C. Burl, U. M. Fayyad, and P. Perona (in press). Knowledge discovery in large image databases: dealing with uncertainties in ground truth. In *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurasamy (eds.), AAAI/MIT Press.

Smyth, P. (1995). Learning with probabilistic supervision. In *Computational Learning Theory and Natural* Learning Systems 3, T. Petsche, S. Hanson, and J. Shavlik, Cambridge, MA: MIT Press, pp.163- 182.

Titterington, D. M. (1989). An alternative stochastic supervisor in discriminant analysis. *Pattern Recognition, vol.22, no.1,* pp.91-95.

Uebersax, J. S. (1 993). Statistical modeling of expert ratings on medical treatment appropriateness. *J. Amer. Statist. Assoc.,* vol.88, no.422, pp.421- *427.*